

UNITED STATES PATENT APPLICATION FOR:

**METHOD AND APPARATUS FOR UTILITY-BASED DYNAMIC RESOURCE
ALLOCATION IN A DISTRIBUTED COMPUTING SYSTEM**

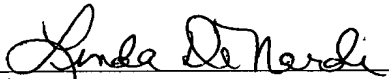
INVENTORS:

**RAJARSHI DAS
JEFFREY O. KEPHART
GERALD J. TESAURO
WILLIAM E. WALSH**

ATTORNEY DOCKET NUMBER: YOR920040007US1

CERTIFICATION OF MAILING UNDER 37 C.F.R. 1.10

I hereby certify that this New Application and the documents referred to as enclosed therein are being deposited with the United States Postal Service on January 30, 2004, in an envelope marked as "Express Mail United States Postal Service", Mailing Label No. EV 177156979 US, addressed to: Mail Stop PATENT APPLICATION, Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450.


Signature

LINDA DENARDI
Name

January 30, 2004
Date of signature

MOSER, PATTERSON & SHERIDAN, LLP
595 Shrewsbury Avenue
Shrewsbury, New Jersey 07702
(732) 530-9404

METHOD AND APPARATUS FOR UTILITY-BASED DYNAMIC RESOURCE ALLOCATION IN A DISTRIBUTED COMPUTING SYSTEM

BACKGROUND

[0001] The present invention relates generally to data processing systems, and relates more particularly to the management of hardware and software components of data processing systems. Specifically, the present invention provides a method and apparatus for automatic allocation of computing resources amongst multiple entities that obtain value by utilizing the resources to perform computation.

[0002] The problem of how to optimally allocate a limited set of resources amongst multiple entities that use or consume the resources has been extensively studied in disciplines including economics, manufacturing, telecommunications networks, and computing systems. Within the latter domain, the recent evolution of highly interconnected, rapidly changing, distributed computing systems such as the Internet has made it increasingly important to be able to rapidly compute and execute resource allocation decisions in an automated fashion.

[0003] Traditional approaches to provisioning and capacity planning typically aim to achieve an extremal value of some overall system performance metric (*e.g.*, maximum average throughput or minimum average response time). Other conventional techniques employ market-based mechanisms for resource allocation (*e.g.*, auction bidding or bilateral negotiation mechanisms). For example, a commonly used approach has been to anticipate the maximum possible load on the system, and then perform one-time static allocation of resources capable of handling the maximum load within a specified margin of safety. A common problem with such approaches is that, with modern workloads such as hit rates on Web pages, the demand rate may vary dynamically and rapidly over many orders of magnitude, and a system that is statically provisioned for its peak workload may spend nearly all its time sitting idle.

[0004] Thus, there is a need in the art for a method and apparatus for dynamic resource allocation in distributed computing systems.

SUMMARY OF THE INVENTION

[0005] In one embodiment, the present invention is a method for optimal and automatic allocation of finite resources (*e.g.*, hardware or software that can be used within any overall process that performs computation) amongst multiple entities that can provide computational services given the resource(s). One embodiment of the inventive method involves establishing, for each entity, a service level utility indicative of how much business value is obtained for a given level of computational system performance and for a given level of demand for computing service. Each entity is capable of transforming its respective service-level utility into a corresponding resource-level utility indicative of how much business value may be obtained for a given set or amount of resources allocated to the entity. The resource-level utilities for each entity are aggregated, and resource allocations are subsequently determined and executed based upon the dynamic resource-level utility information established. The invention is thereby capable of making rapid allocation decisions, according to time-varying need or value of the resources by each of the entities. In addition, the inventive method is motivated by the perspective of an enterprise comprising multiple entities that use said finite computational resources to provide service to one or more customers, and is thus structured to optimize the business value of the enterprise.

BRIEF DESCRIPTION OF THE DRAWINGS

[0006] So that the manner in which the above recited embodiments of the invention are attained and can be understood in detail, a more particular description of the invention, briefly summarized above, may be obtained by reference to the embodiments thereof which are illustrated in the appended drawings. It is to be noted, however, that the appended drawings illustrate only typical embodiments of this invention and are therefore not to be considered limiting of its scope, for the invention may admit to other equally effective embodiments.

[0007] Figure 1 is a diagram of a networked data processing system in which the present invention may be implemented;

[0008] Figure 2 is an overall view of a resource allocation system in accordance with one embodiment of the present invention;

[0009] Figure 3 is a flow chart illustrating one embodiment of a method for dynamically allocating resources among multiple application environments;

[0010] Figure 4 is a diagram illustrating the detailed functionality of an application environment module which constitutes a component of the overall system shown in Figure 2; and

[0011] Figure 5 is a high level block diagram of the present invention implemented using a general purpose computing device.

[0012] To facilitate understanding, identical reference numerals have been used, where possible, to designate identical elements that are common to the figures.

DETAILED DESCRIPTION

[0013] In one embodiment, the present invention is a method for optimal and automatic allocation of finite resources amongst multiple entities that can perform computational work given the resource(s). For the purposes of the present invention, the term “resource” may indicate an entire hardware or software component (*e.g.*, a compute server, a storage device, a RAM circuit or a database server), or a portion of a component (*e.g.*, bandwidth access or a fraction of a server). The method may be implemented, for example, within a data processing system such as a network, a server, or a client computer. The invention is capable of making allocation decisions in real time, according to time-varying need or value of the resources by each of the entities, thereby resolving the shortcomings associated with typical static resource allocation techniques. In addition, the method is structured to optimize the business value of an enterprise that provides computing services to multiple entities using said finite computational resources.

[0014] Figure 1 is a schematic illustration of one embodiment of a network data processing system 100 comprising a network of computers (*e.g.*, clients) in which the present invention may be implemented. The network data processing system 100

includes a network 102, a server 104, a storage unit 106 and a plurality of clients 108, 110 and 112. The network 102 is the medium used to provide communications links between the server 104, storage unit 106 and clients 108, 110, 112 connected together within network data processing system 100. The network 102 may include connections, such as wire, wireless communication links, or fiber optic cables.

[0015] In the embodiment illustrated, the server 104 provides data, such as boot files, operating system images, and applications to the clients 108, 110, 112 (*i.e.*, the clients 108, 110, and 112 are clients to server 104). The clients 108, 110, and 112 may be, for example, personal computers or network computers. Although the network data processing system 100 depicted in Figure 1 comprises a single server 104 and three clients, 108, 110, 112, those skilled in the art will recognize that the network data processing system 100 may include additional servers, clients, and other devices not shown in Figure 1.

[0016] In one embodiment, the network data processing system 100 is the Internet, with the network 102 representing a worldwide collection of networks and gateways that use the Transmission Control Protocol/Internet Protocol (TCP/IP) suite of protocols to communicate with one another. In further embodiments, the network data processing system 100 is implemented as an intranet, a local area network (LAN), or a wide area network (WAN). Furthermore, although Figure 1 illustrates a network data processing system 100 in which the method of the present invention may be implemented, those skilled in the art will realize that the present invention may be implemented in a variety of other data processing systems, including servers (*e.g.*, server 104) and client computers (*e.g.*, clients 108, 110, 112). Thus, Figure 1 is intended as an example, and not as an architectural limitation for the present invention.

[0017] Figure 2 is a schematic illustration of one embodiment of a data center 200 for executing the method of the present invention. The data center 200 comprises a plurality of application environment modules 201, 202, and 203, one or more resource arbiters 204 and a plurality of resources 205, 206, 207, 208 and 209. Each application environment module 201-203 is responsible for handling respective demands 213,

214 and 215 (*e.g.*, requests for information processing services) that may arrive from a particular customer or set of clients (*e.g.*, clients 108-112 in Figure 1). Example client types include: online shopping services, online trading services, and online auction services.

[0018] In order to process client demands 213, 214 or 215, the application environments 201-203 may utilize the resources 205-209 within the data center 200. As each application environment 201-203 is independent from the others and provides different services, each application environment 201-203 has its own set of resources 205-209 at its disposal, the use of which must be optimized to maintain the appropriate quality of service (QoS) level for the application environment's clients. An arrow from an application environment 201-203 to a resource 205-209 denotes that the resource 205-209 is currently in use by the application environment 201-203 (*e.g.*, in Figure 2, resource 205 is currently in use by application environment 201). An application environment 201-203 also makes use of data or software objects, such as respective Service Level Agreements (SLAs) 210, 211 and 212 with its clients, in order to determine its service-level utility function $U(S,D)$. An example SLA 210-212 may specify payments to be made by the client based on mean end-to-end response time averaged over, say, a five-minute time interval. Additionally the client workload may be divided into a number of service classes (*e.g.*, Gold, Silver and Bronze), and the SLA 210-212 may specify payments based on details of response time characteristics within each service class.

[0019] Each application environment 201-203 is in further communication with the resource arbiter module 204. Although the data center 200 illustrated in Figure 2 utilizes only one resource arbiter 204, those skilled in the art will appreciate that multiple resource arbiters may be implemented in the data center 200. The resource arbiter 204 is responsible for deciding, at any given time while the data center 200 is in operation, which resources 205-209 may be used by which application environments 201-203. In one embodiment, the application environments 201-203 and resource arbiter 204 are software modules consisting of autonomic elements (*e.g.*, software components that couple conventional computing functionality with additional self-management capabilities), for example written in Java™, and

YOR920040007US1

communication between modules 201-203 and 204 takes place using standard Java interfaces. The modules 201-203 and 204 may run on a single computer or on different computers connected by a network such as the Internet or a Local Area Network (LAN), *e.g.*, as depicted in Figure 1. In the networked case, communication may additionally employ standard network communication protocols such as TCP/IP and HTTP, and standard Web interfaces such as OGSA.

[0020] Figure 3 is a flow chart illustrating the method 300 by which the resource arbiter 204 makes resource allocation decisions. Referring simultaneously to Figures 2 and 3, the method 300 is initialized at block 302 and proceeds to block 304, where the method 300 establishes a service-level utility function $U(S, D)$ for each application environment 201-203. In one embodiment, the variable S is a vector that characterizes the multiple performance measures for multiple service classes, and the variable D is a vector that characterizes the demand. The service level utility indicates how much business value U is obtained by the application environment 201, 202 or 203 for various levels S of computational system performance, and for a given level D of demand 213-215 for computing service.

[0021] In one embodiment, the service-level utility function $U(S, D)$ is established by the application environment's SLA 210-212. While each application environment's service-level utility may be based on different performance metrics, all of the service-level utility functions $U(S, D)$ share a common scale of valuation.

[0022] In block 306, the method 300 transforms the service-level utility function $U(S, D)$ into a resource-level utility function $V(R)$ for each application environment 201-203. The resource level utility indicates how much business value V is obtained for a given actual or hypothetical set or amount of resources R (*e.g.*, selected from resources 205-209) allocated to the application environment 201-203. In one embodiment, R is a vector. For example, the utility information may express a utility curve $V(m)$, the utility obtained from being able to use m compute servers, at various values of m ranging from 0 to the total number of compute servers within the data center. Additionally if the servers are of different types, the utility information may express the value of obtaining m servers of type A, n servers of type B, etc. More

generally the utility information may express $V(\{x\})$, the value of assigning a particular collection or set $\{x\}$ of resources 205-209, for various sets $\{x\}$ ranging over the power set of possible resources 205-209 that could be assigned to the application environment 201-203. The utility information may be expressed, for example, in a parameterized functional form, or it may also be expressed in terms of values at a set of discrete points which may represent a subset or complete set of all possible resource levels that could be provided.

[0023] The transformation may additionally depend on a set of variables describing the application environment's current state (*e.g.*, current demand 213-215, system load, throughput or average response time), or on differences between a hypothetical resource allocation R and the application environment's current resource allocation R^* (*e.g.*, in a manner that reflects any costs associated with switching the allocation from R^* to R , including delays, machine downtime, etc.). In one embodiment, the resource-level utility function is calculated according to the relation

$$V_i(R_i) = U_i(S_i, D_i, R_i) \quad (\text{EQN. 1})$$

such that $S_i \in S_i(R_i, D_i)$, where $S_i(R_i, D_i)$ is a relation specifying the set of service levels attainable with resources R_i and demand D_i . In one embodiment, the relation $S_i(R_i, D_i)$ is obtained by standard computer systems modeling techniques (*e.g.*, queuing theory). In another embodiment, the relation $S_i(R_i, D_i)$ may instead or additionally be refined by training on a collection of observed system performance data $\{(S_i, R_i, D_i)\}$ using standard machine learning procedures (*e.g.*, supervised learning methods employing standard linear or nonlinear function approximators).

[0024] In one embodiment, the resource-level utility function $V(R)$ estimates the current value of the current state. In another embodiment, the resource-level utility function estimates the expected cumulative discounted or undiscounted future value starting from the current state. In one embodiment, any one or more of a number of standard methodologies may be employed in the process of estimating expected future value, including prediction and forecasting methodologies such as time-series prediction methods and machine learning methodologies such as reinforcement

learning algorithms (*e.g.*, Q-Learning, Temporal Difference Learning, R-Learning or SARSA).

[0025] In block 308, the method 300 communicates the respective resource-level utility functions for each application environment 201-203 to the resource arbiter 204 and aggregates all resource level utility functions. In one embodiment, while the data center 200 is running, from time to time each application environment 201-203 communicates to the resource arbiter 204 information regarding its current resource-level utility function. Said communication may take place either synchronously or asynchronously, and may be initiated by the application environments 201-203, or may be in response to a prompt or query issued by the resource arbiter 204.

[0026] In block 310, the method 300, having received resource-level utility information from each application environment 201-203, combines said utility information and thereupon decides how to assign each available resource 205-209 in the data center 200, in a manner that optimizes the total utility obtained. In other words, the resource arbiter 204 maximizes the sum of the resource-level utilities,

$\max_{R \in R} \sum_i V_i(R_i)$. Said resource assignment may include the possibility of a null

assignment, (*i.e.*, the resource 205-209 is not assigned to any application environment 201-203) so that the resource 205-209 may be kept in reserve to handle future workload. For example, in the case of undifferentiated compute servers within the data center 200, the resource arbiter 204 may utilize the most recent utility curves from each application environment 201-203 ($V_1(m)$, $V_2(m)$ and $V_3(m)$ respectively), and then compute an integral number of servers (m_1 , m_2 , m_3) to assign to each application environment 201-203 so as to maximize the total $V_1(m_1) + V_2(m_2) + V_3(m_3)$. The determination of an allocation that optimizes total utility will generally be made by executing an optimization method. In one embodiment, the values (m_1 , m_2 , m_3) are found by using standard linear or nonlinear algorithms such as hill climbing, simulated annealing, linear programming, or mixed-integer programming. Additionally, the objective function optimized by the resource arbiter 204 may also include any switching costs that are incurred when a particular resource 205-209 is reallocated from one application environment 201-203 to another. Said switching

costs may include, for example, machine downtime and/or other costs related to installing or removing data or software from the machine when it is reallocated.

[0027] In block 312, the method 300 executes the resource allocation decision calculated in block 310, and communicates the resource allocation decision to the application environments 201-203. In one embodiment, block 312 additionally involves the causation of manipulations or operations performed upon the resources 205-209, enabling the resources 205-209 to be used by the application environments 201-203 to which the resources 205-209 have been assigned, or associated with de-allocating a resource 205-209 from an application environment 201-203 to which the resource 205-209 is no longer assigned.

[0028] Figure 4 is a schematic illustration of the basic operations and functionality of one embodiment of an application environment module 401 according to the present invention, wherein the application environment module 401 is any of the application environments 201-203 depicted in Figure 2. In one embodiment, the application environment module 401 comprises an autonomic manager element 402, a workload router 403, and a system performance monitoring element 404. Interactions of the application environment 401 with its SLA 410, its client demand 411, its currently allocated resources (*e.g.*, compute servers 420, 421, and 422), and with the resource arbiter element 412, are depicted as they were in Figure 2.

[0029] While the application environment 401 is in operation, from time to time client demand 411 is received and transmitted to the router 403, which thereupon sends said demand 411 to one of the assigned compute servers 420, 421, or 422, typically based on the use of a routing or load-balancing method. As client jobs are processed, their intermediate and final output are returned to the submitting client. From time to time the performance monitor 404 may observe, request or receive information regarding measures or statistics of the system performance of the compute servers 420-422, such as CPU/memory usage, average throughput, average response time, and average queue depth. The autonomic manager 402 combines said performance measures with information regarding the demand 411, the SLA 610, and

the currently allocated resources 420-422, to produce an estimated resource-level utility function.

[0030] In one embodiment, said utility function indicates $V(m)$, the value of being allocated an integral quantity m of undifferentiated compute servers, with the value of m ranging from zero to the total number of servers in the data center (*e.g.*, data center 200 in Figure 2). From time to time said utility function is transmitted to the resource arbiter 412, possibly in response to a prompt or query sent from the resource arbiter 412. From time to time said resource arbiter 412 will additionally transmit to the application environment 401 updated information regarding its set of allocated resources. The updated information indicates, for example, that certain compute servers 420-422 are newly available for usage, or that certain compute servers 420-422 previously used by the application environment 401 are to be de-allocated and are no longer available for usage.

[0031] In another embodiment, the autonomic manager module 402 of Figure 4 further comprises a capability to model the effect of any adjustable operational parameters the resources 420-422 may have (*e.g.*, maximum queue depth, buffer pool sizes, etc.) on the observed system performance. The autonomic manager 402 further operates to set said parameters of the resources 420-422, or of the router 403, or other internal parameters, to values such that the resulting system-level utility function optimizes the resource-level utility function.

[0032] In another embodiment of the invention, the autonomic manager module 402 of Figure 4 further comprises a capability to model or predict the demand at future times given the observed current demand 411, and a capability to model or predict the system performance at future times given the current demand 411, current performance, and future allocated resources, which may be the same or different from the current allocated resources 420-422. The autonomic manager 402 then computes a resource-level utility function indicating the cumulative discounted or undiscounted future utility associated with a hypothetical resource allocation made at the current time. In one embodiment, the predicted demand and predicted system performance are deterministic predictions at each future time. In another embodiment, the

predicted demand and predicted system performance are probability distributions over possible levels of demand or performance at each future time. In one embodiment, the cumulative future utility is obtained by summation over a finite number of discrete future time steps. In another embodiment, the cumulative future utility is obtained by integration over a continuous future time interval.

[0033] In another embodiment of the invention, the autonomic manager module 402 of Figure 4 does not explicitly predict future demand or future system performance, but instead uses machine learning procedures to estimate cumulative discounted or undiscounted future utility from a temporal sequence of observed data points, each data point consisting of: an observed demand, an observed system performance, an observed resource allocation, and an observed payment as specified by the SLA 410. In one embodiment, the machine learning procedure consists of a standard reinforcement learning procedure such as Q-Learning, Temporal Difference Learning, R-Learning or SARSA.

[0034] Figure 5 is a high level block diagram of the present dynamic resource allocation system that is implemented using a general purpose computing device 500. In one embodiment, a general purpose computing device 500 comprises a processor 502, a memory 504, a dynamic resource allocator or module 505 and various input/output (I/O) devices 506 such as a display, a keyboard, a mouse, a modem, and the like. In one embodiment, at least one I/O device is a storage device (*e.g.*, a disk drive, an optical disk drive, a floppy disk drive). It should be understood that the dynamic resource allocator 505 can be implemented as a physical device or subsystem that is coupled to a processor through a communication channel.

[0035] Alternatively, the dynamic resource allocator 505 can be represented by one or more software applications (or even a combination of software and hardware, *e.g.*, using Application Specific Integrated Circuits (ASIC)), where the software is loaded from a storage medium (*e.g.*, I/O devices 506) and operated by the processor 502 in the memory 504 of the general purpose computing device 500. Thus, in one embodiment, the resource allocator 505 for allocating resources among entities described herein with reference to the preceding Figures can be stored on a computer

readable medium or carrier (*e.g.*, RAM, magnetic or optical drive or diskette, and the like).

[0036] The functionalities of the arbiters and the application environments described with reference to Figures 2 and 4 may be performed by software modules of various types. For example, in one embodiment, the arbiters and/or application environments comprise autonomic elements. In another embodiment, the arbiters and/or application environments comprise autonomous agents software as may be constructed, for example, using the Agent Building and Learning Environment (ABLE). The arbiters and/or application environments may all run on a single computer, or they may run independently on different computers. Communication between the arbiters and the application environments may take place using standard interfaces and communication protocols. In the case of arbiters and application environments running on different computers, standard network interfaces and communication protocols may be employed, such as Web Services interfaces (*e.g.*, those employed in the Open Grid Services Architecture (OGSA)).

[0037] Thus, the present invention represents a significant advancement in the field of dynamic resource allocation. A method and apparatus are provided that enable a finite number of resources to be dynamically allocated among a number of entities or application environments capable of performing computational work given the resources. The allocation is performed in a manner that optimizes the business value of the enterprise providing the computing services to a number of clients.

[0038] While foregoing is directed to the preferred embodiment of the present invention, other and further embodiments of the invention may be devised without departing from the basic scope thereof, and the scope thereof is determined by the claims that follow.